Emerging domain agnostic functionalities on the handle-centered networks

Kei Kurakawa[1*], ***Takayuki Sekiya***[2], ***Yasumasa Baba***[3]

[1*] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, 101-8430, Japan*
[2] *The University of Tokyo, 3-8-1 Komaba, Meguro, Tokyo, 153-8902, Japan*
[3] *The Institute of Statistical Mathematics, 10-3 Midoricho, Tachikawa, Tokyo, 190-8562, Japan*
Email: kurakawa@nii.ac.jp

Summary. Core part of RDA WGs and IGs aims at establishing discovery and automatic processing way of data. This work illustrates automatic data processing framework in detail and implies an emerging domain independent functionalities on the framework.

Keywords. Digital Object, Handle, Automatic data processing, Kernel information, Domain independent functionalities.

1. Introduction

The infrastructure for scientists to share the scientific resources such as journal articles, conference proceedings, books, software, data, and any other available online resources has been developed over several decades. Recent trend of Open Access movement since 2000s reformulate the scientific community's mind so as to make the scientific resources open as default to be freely accessible by everyone. The scientific community's mind of Open Access has influenced on the divergence range of academic resources to share, as a result the movement reached at the slogan "research data sharing without barriers" of RDA (Research Data Alliance) among all disciplines.

Our standard procedures, which may be peculiar to each discipline, to aggregate and process the scientific data are a process of craftsmanship and a too much time consuming task compared to academic rewards, in that to deal with scientific data the data consumer needs to understand the semantics of data structure in domain dependent schemes and choose ordinarily a community standard of tools on a specific computational environment to process the data, which seems to be difficult to do the same things by outsiders of the expertise. The whole kind of the things must be easy and automatic even for the experts, needless to say for all who need the data.

Core part of RDA working and interest groups (WGs and IGs) concentrate on the issues and tackle with the PID (persistent identifiers) centric information model to invest the domain-independent automatic processing environment for very large and heterogeneous collections of distributed scientific data. This paper introduces the relevant information model produced by the WGs and IGs, and considers the emerging domain independent functionalities, which can be derived from the information model.

2. Data discovery and automatic data processing

To think of a next generation of research data infrastructure we need to analyze research data processing workflows of researchers. The major requirements are two kinds as follows.

One is to enhance discoverability of scientific research data that are explicitly separated from general resources on the Web. RDA Data Discovery IG focuses on the discoverable search interfaces and technical specific metadata for the scientific research data. For geophysical science a typical user interface has a browsing feature with longitude and latitude settings on geographical coordinate, which requires a metadata

enhancement from a common Dublin Core metadata scheme. Domain specific metadata schemes for a certain kind of disciplines may deductively produce available interfaces for search, but they are difficult to unify among a variety of scientific domains.

The other is to build automatic processing environment for scientific research data. Several groups of RDA (Data Fabric, Data Type Registries, PID Information Types, and PID Kernel Information) focus on this vision, especially in domain independent way, and discuss what the necessary information framework is for the future research environment.

3. Automatic data processing framework

Automatic data processing is an ideal and urgent way of function to deal with distributed large amount of data. The followings are the technological views to implement the framework.

3.1 Digital object and data types

Persistent identifiers (PIDs) as known as handles is a key to point out permanently a specific dataset on the network. The Handle Server is the implementation based on Digital Object Architecture [1] also known as Kahn-Wilensky Framework issued in 1995, which have influenced several digital library architecture development for a number of decades [2]. A Digital Object (DO) consists of data and key-metadata. The key-metadata includes a globally unique PID namely a handle and other metadata. In the architecture, the data of each digital object is associated with a type stored in a global data type registry.

3.2 Versioning and provenance information

To state the quality and context of data in general, versioning and provenance information used to be embedded in the metadata. Provenance information gives rich context of data such as quality, audit trail, replication, attribution, etc [3].

3.3 Domain dependent and independent metadata layers

Metadata consists of a variety of attributes pertaining to the data. Attributes of data are defined and cross-related in different granularity of concept, in different levels of domain-specificity, and in different aspects. Domain independent attributes are collected in Kernel Information profile as structural metadata [4].

3.4 Handle-centered networks

All kinds of things in this framework, e.g., data, types, metadata are embedded with a handle as a pointer. Metadata describes relationships among them by attributes to a handle. The collection of metadata produces a handle-centered network.

4. Domain independent functionalities on the framework

The handle-centered networks can be source for analytics to produce valuable knowledge, i.e.,

- Analysis of data
- Classification of data
- Recommendation of data
- Prediction of data.

5. Conclusions

This work illustrated the automatic data processing framework discussed in RDA community, and implied emerging domain independent functionalities on the framework.

Acknowledgments. This work is supported by the open collaborative research at National Institute of Informatics (NII) Japan (FY2017). The authors are thankful to all RDA Kernel Information WG members for their great discussions on remotely and in-person meetings.

References

1. Kahn, R., Wilensky, R., A Framework for Distributed Digital Object Services, doi: cnri.dlib/tn95-01, 1995
2. Nelson, M. L., Sompel, H. Van de, IJDL special issue on complex digital objects: Guest editors' introduction. *Int. J. Digit. Libr.,* 6,113-114, doi: 10.1007/s00799-005-0127-y, 2006
3. Simmhan, Y. L., Plale, B., Gannon, D., A survey of data provenance in e-science. *ACM SIGMOD Record,* 34(3), 31, doi:10.1145/1084805.1084812, 2005
4. RDA KI WG, Strawman PID Kernel Information Profile 17.04.05., http://bit.ly/2oH53XC, [accessed on October, 2017]